Лекция 8. Принцип максимального сходства и логистическая регрессия. L2 регулиризация

Принцип максимального правдоподобия и логистическая регрессия

- Теперь посмотрим, как из принципа максимального правдоподобия получается оптимизационная задача, которую решает логистическая регрессия, а именно, – минимизация *логистической* функции потерь.

Принцип максимального правдоподобия и логистическая регрессия

Теперь посмотрим, как из принципа максимального правдоподобия получается оптимизационная задача, которую решает логистическая регрессия, а именно, – минимизация *логистической* функции потерь.
Только что мы увидели, что логистическая регрессия моделирует вероятность отнесения примера к классу "+" как

$$p_+(\overrightarrow{x_i}) = P\left(y_i = 1 \mid \overrightarrow{x_i}, \vec{w}\right) = \sigma(\vec{w}^T \overrightarrow{x_i})$$

Тогда для класса "-" аналогичная вероятность:

$$p_-(\overrightarrow{x_i}) = P\left(y_i = -1 \mid \overrightarrow{x_i}, \vec{w}\right) = 1 - \sigma(\vec{w}^T \overrightarrow{x_i}) = \sigma(-\vec{w}^T \overrightarrow{x_i})$$

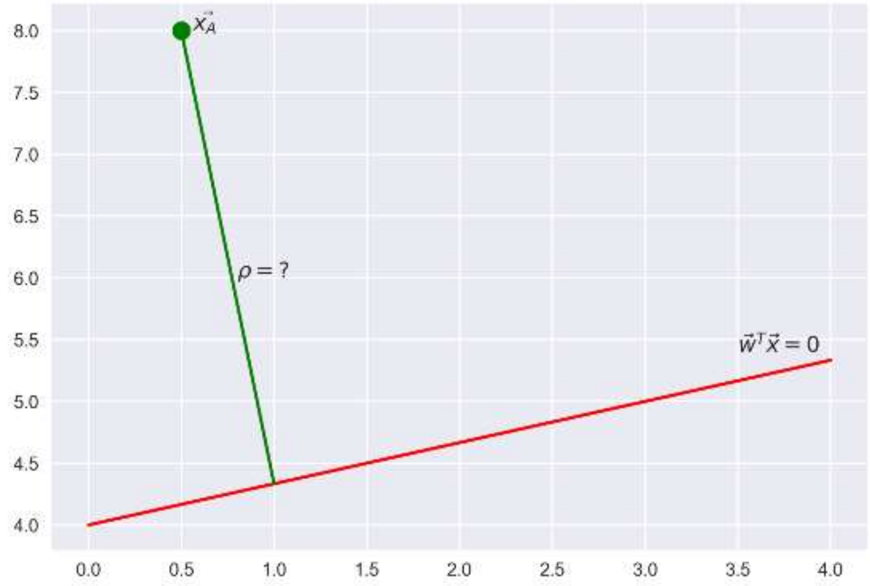
Оба этих выражения можно ловко объединить в одно (следите за моими руками – не обманывают ли вас):

$$P\left(y = y_i \mid \overrightarrow{x_i}, \vec{w}\right) = \sigma(y_i \vec{w}^T \overrightarrow{x_i})$$

Выражение $M(\overrightarrow{x_i}) = y_i \vec{w}^T \overrightarrow{x_i}$ называется *отступом* (*margin*) классификации на объекте $\overrightarrow{x_i}$ (не путать с зазором (тоже margin), про который чаще всего говорят в контексте SVM). Если он неотрицателен, модель не ошибается на объекте $\overrightarrow{x_i}$, если же отрицателен – значит, класс для $\overrightarrow{x_i}$ спрогнозирован неправильно.

Заметим, что отступ определен для объектов именно обучающей выборки, для которых известны реальные метки целевого класса y_i.

$$\rho(\overrightarrow{x_A}, \vec{w}^T \vec{x} = 0) = \frac{\vec{w}^T \overrightarrow{x_A}}{||\vec{w}||}$$

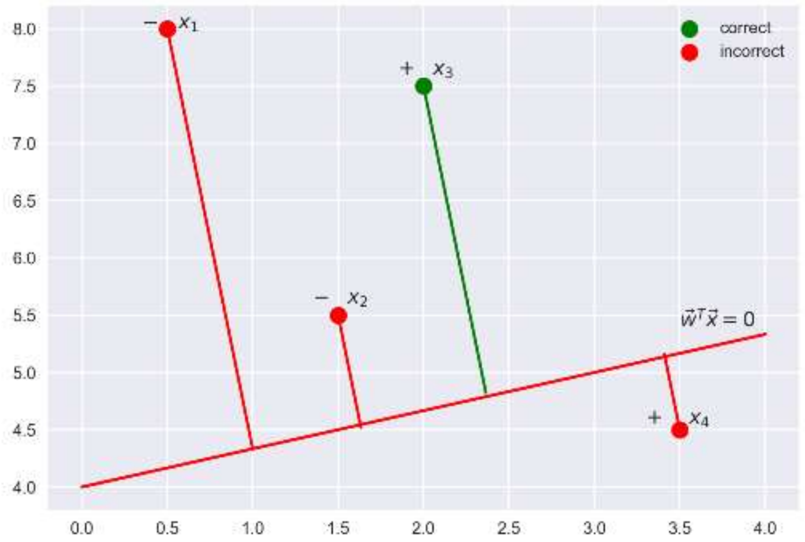


Когда получим (или посмотрим) ответ, то поймем, что чем больше по модулю выражение $\vec{w}^T \overrightarrow{x_i}$, тем дальше точка $\overrightarrow{x_i}$ находится от плоскости $\vec{w}^T \vec{x} = 0$.

Значит, выражение $M(\overrightarrow{x_i}) = y_i \vec{w}^T \overrightarrow{x_i}$ – это своего рода "уверенность" модели в классификации объекта $\overrightarrow{x_i}$:

Значит, выражение $M(\overrightarrow{x_i}) = y_i \vec{w}^T \overrightarrow{x_i}$ – это своего рода "уверенность" модели в классификации объекта $\overrightarrow{x_i}$:

- если отступ большой (по модулю) и положительный, это значит, что метка класса поставлена правильно, а объект находится далеко от разделяющей гиперплоскости (такой объект классифицируется уверенно). На рисунке – x_3.
- если отступ большой (по модулю) и отрицательный, значит метка класса поставлена неправильно а объект находится далеко от разделяющей гиперплоскости (скорее всего такой объект – аномалия, например, его метка в обучающей выборке поставлена неправильно). На рисунке – x_1.
- если отступ малый (по модулю), то объект находится близко к разделяющей гиперплоскости, а зна отступа определяет, правильно ли объект классифицирован. На рисунке – x_2 и x_4.



Теперь распишем правдоподобие выборки, а именно, вероятность наблюдать данный вектор $\vec{y}$ у выборки X. Делаем сильное предположение: объекты приходят независимо, из одного распределения (*i.i.d.*). Тогда

$$P(\vec{y} \mid X, \vec{w}) = \prod_{i=1}^{\ell} P\left(y = y_i \mid \overrightarrow{x_i}, \vec{w}\right),$$

где ℓ – длина выборки X (число строк).

Как водится, возьмем логарифм данного выражения (сумму оптимизировать намного проще, чем произведение):

$$\begin{aligned} \log P(\vec{y} \mid X, \vec{w}) &= \log \prod_{i=1}^{\ell} P\left(y = y_i \mid \overrightarrow{x_i}, \vec{w}\right) \\ &= \log \prod_{i=1}^{\ell} \sigma(y_i \vec{w}^T \overrightarrow{x_i}) \\ &= \sum_{i=1}^{\ell} \log \sigma(y_i \vec{w}^T \overrightarrow{x_i}) \\ &= \sum_{i=1}^{\ell} \log \frac{1}{1+\exp^{-y_i \vec{w}^T \overrightarrow{x_i}}} \\ &= -\sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \overrightarrow{x_i}}) \end{aligned}$$

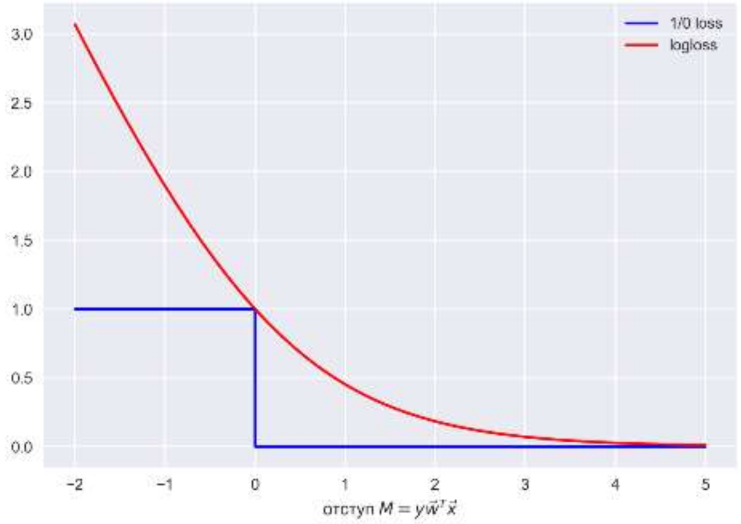
То есть в даном случае принцип максимизации правдоподобия приводит к минимизации выражения

То есть в даном случае принцип максимизации правдоподобия приводит к минимизации выражения

$$\mathcal{L}_{log}(X, \vec{y}, \vec{w}) = \sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \vec{x_i}}).$$

Это *логистическая* функция потерь, просуммированная по всем объектам обучающей выборки.

Посмотрим на новую фунцию как на функцию от отступа: $L(M) = \log(1 + \exp^{-M})$. Нарисуем ее график, а также график 1/0 функциий потерь (*zero-one loss*), которая просто штрафует модель на 1 за ошибку на каждом объекте (отступ отрицательный): $L_{1/0}(M) = [M < 0]$.



Картинка отражает общую идею, что в задаче классификации, не умея напрямую минимизировать число ошибок (по крайней мере, градиентными методами это не сделать – производная 1/0 функциий потерь в нуле обращается в бесконечность), мы минимизируем некоторую ее верхнюю оценку. В данном случае это логистическая функция потерь (где логарифм двоичный, но это не принципиально), и справедливо

$$\begin{aligned}\mathcal{L}_{1/0}(X, \vec{y}, \vec{w}) &= \sum_{i=1}^{\ell} [M(\overrightarrow{x_i}) < 0] \\ &\leq \sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \overrightarrow{x_i}}) \\ &= \mathcal{L}_{log}(X, \vec{y}, \vec{w})\end{aligned}$$

где $\mathcal{L}_{1/0}(X, \vec{y}, \vec{w})$ – попросту число ошибок логистической регрессии с весами $\vec{w}$ на выборке $(X, \vec{y})$.

То есть уменьшая верхнюю оценку $\mathcal{L}_{log}$ на число ошибок классификации, мы таким образом надеемся уменьшить и само число ошибок.

L_2-регуляризация логистических потерь

L2-регуляризация логистической регрессии устроена почти так же, как и в случае с гребневой (Ridge регрессией). Вместо функционала $\mathcal{L}_{log}(X, \vec{y}, \vec{w})$ минимизируется следующий:

$$J(X, \vec{y}, \vec{w}) = \mathcal{L}_{log}(X, \vec{y}, \vec{w}) + \lambda |\vec{w}|^2$$

В случае логистической регрессии принято введение обратного коэффициента регуляризации $C = \frac{1}{\lambda}$.